AD-A096 952 STANFORD UNIV CA DEPT OF STATISTICS F/G 12/1
NEAREST NEIGHBOR CLASSIFICATION OF STATIONARY TIME SERIES: AN A--ETC(U)
DEC 80 W GERSCH N00014-76-C-0475
UNCLASSIFIED TR-294 NL

1 or 1
AD
A096952

END
DATE
FILMED
4-81
DTIC

LEVEL II

14

NEAREST NEIGHBOR CLASSIFICATION OF STATIONARY TIME SERIES: AN APPLICATION TO ANESTHESIA LEVEL CLASSIFICATION BY EEG ANALYSIS

By

10 WILL/GERSCH

9 TECHNICAL REPORT NO. 294

14 TR-294

11 5 DEC[redacted]80

12

Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATSITICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

NEAREST NEIGHBOR CLASSIFICATION OF STATIONARY TIME SERIES: AN APPLICATION TO ANESTHESIA LEVEL CLASSIFICATION BY EEG ANALYSIS

By
Will Gersch

1. INTRODUCTION.

This paper presents the theory and a prototypic example of an exploratory population screening-stationary time series classification problem. In the population screening problem a new individual is classified by comparing measurements obtained from him with measurements obtained from other individuals in the alternative categorical states. Human electroencephalogram (EEG) time series were obtained during surgery simultaneously with an anesthesiologist's appraisal of the level of anesthesia from a moderate but not large number of individuals. These EEG time series are considered to be a set of labeled sample time series. The categorical time series classes are characterized by broad intersubject time series variations. An implicit conjecture in this data gathering experiment is that there is sufficient information in the EEG time series to reliably classify the level of anesthesia of humans in surgery. Our objectives are to assess the separability of the time series populations, i.e. to obtain a statistically reliable estimate of the minimum achievable probability of misclassification of new time series and to implement a time series classification rule that can achieve those statistical properties.

A nearest neighbor time series classification rule achieves those objectives. With that rule a measure of dissimilarity is computed between a new to-be-classified time series and each of a set of categorically labeled time series. The new time series is classified with the label of its least dissimilar neighbor. In our approach the dissimilarity measure between time series is an estimate of the Kullback Leibler number between the time series as if the time series were normally distributed. This dissimilarity measure is shown to have sufficient metric properties for the formal Cover and Hart 1967 asymptotic nearest neighbor and Rogers 1977 finite sample nearest neighbor classification rule properties to hold. Those properties allow the conjecture, that there is sufficient information in the EEG time series to reliably classify the level of anesthesia of humans in surgery, to be tested with only a moderate number of labeled sample EEG time series.

The nearest neighbor Kullback Leibler type dissimilarity measure classification rule (NN-KL) method is applied to the classification of the level of anesthesia of humans in surgery by the analysis of multichannel EEGs. Application of the method exploits time domain formulas for the Kullback Leibler number between multivariate stationary Gaussian time series.

Section 2 describes the nearest neighbor time series classification rule with Kullback Leibler type dissimilarity measure. An implementation and interpretation of the nearest neighbor Kullback Leibler classification rule for the classification of stationary time series is in Section 3. Also in that section, a careful distinction is made between our own use of nearest neighbor Kullback Leibler type dissimilarity classification rules and similarly designated feature analysis-discriminant analysis classification procedures that are common in speech processing. The anesthesia level classification by EEG time series population screening problem example is in Section 4. An appendix shows both the metric properties of the Kullback Leibler type dissimilarity measure and time and frequency domain Kullback Leibler number formulas for multivariate stationary Gaussian time series. Relatively nontechnical discussions of the problem discussed in this paper appear in Gersch et al 1979 and Gersch et al 1980.

2. NEAREST NEIGHBOR RULE CLASSIFICATION WITH A KULLBACK LIEBLER TYPE DISSIMILARITY MEASURE.

Let the labeled sample time series be

$$\begin{bmatrix} x^{(1)} \\ \theta^{(1)} \end{bmatrix}, \ldots, \begin{bmatrix} x^{(N)} \\ \theta^{(N)} \end{bmatrix} \tag{2.1}$$

$$x^{(m)} = (x^{(m)}(1),\ldots,x^{(m)}(T)),\; x^{(m)}(t) = (x_1^{(m)}(t),\ldots,x_d^{(m)}(t))'$$

$$\theta^{(m)} \varepsilon \{1,\ldots,M\} \cdot$$

In equation (2.1) $x^{(m)}$ denotes a d variable - T duration time series, the ' denotes the matrix transpose and $\theta^{(m)}$ denotes the label or category of the m-th time series. There are M alternative categories.

Designate a new to-be-classified time series

$$x^{(o)} = (x^{(o)}(1),\ldots,x^{(o)}(T)) \; . \tag{2.2}$$

The nearest neighbor classification rule is: Let $d(x^{(o)},x^{(m)})$ be a measure of dissimilarity between the new time series $x^{(o)}$ and the labeled time series $x^{(m)}$, for $m = 1,\ldots,N$.

$$\begin{aligned} &\text{If:} \quad d(x^{(o)},x^{(m')}) \le d(x^{(o)},x^{(m)}) \quad m = 1,\ldots,N \\ &\text{Then:} \; \theta^{(o)} = \theta^{(m')} \end{aligned} \tag{2.3}$$

That is, the new series $x^{(o)}$ is given the label of its nearest dissimilarity measure time series.

The dissimilarity measure between time series that we employ for classification is an estimate of the Kullback Leibler number or I-divergence between time series, computed as if the time series were Gaussian distributed. Let X_o and X_m be two d-vector random variabes with probability density functions f_o and f_m respectively. Then, the I-divergence between f_o and f_m is, Kullback, 1968

$$I(f_o,f_m) = \int f_o(x) \log \frac{f_o(x)}{f_m(x)} dx \ . \tag{2.4}$$

In particular, let $X_o \sim \mathcal{N}(\mu_o,\Sigma_o)$ and $X_m \sim \mathcal{N}(\mu_m,\Sigma_m)$. That is let X_o and X_m each be normally distributed with d-component zero-mean vectors and $d \times d$ covariance matrices Σ_o,Σ_m respectively. In that case, from Kullback 1968

$$2\ I(f_o,f_m) = \log \frac{|\Sigma_m|}{|\Sigma_o|} + \mathrm{tr}\Sigma_m{}^{-1}\Sigma_o - d \quad . \tag{2.5}$$

In equation (2.5) and subsequently, the notation $|A|$, $\mathrm{tr}(A), A^{-1}$. denotes respectively the determinant, trace, inverse of the matrix A.

Consider the d variate-T duration labeled sample time series $x^{(m)}$ $m = 1,\ldots,N$ and the new time series $x^{(o)}$. Let $\hat{\Sigma}_j$, $j = o$ or m be the sample or estimated covariance matrices respectively of $x^{(j)}$ with $\hat{\Sigma}_j = [\hat{\gamma}_{r,c}]$; and $\hat{\gamma}_{r,c}$, the r-c row-column element of $\hat{\Sigma}_j$. Then, let

$$d(x^{(o)},x^{(m)}) = \frac{1}{2Td}\left[\ln \frac{|\hat{\Sigma}_m|}{|\hat{\Sigma}_m|} + \mathrm{tr}\ \Sigma_m^{-1}\Sigma_o - dT \right] \tag{2.6}$$

denote a measure of the dissimilarity computed between the sample time series $x^{(o)}$ and $x^{(m)}$. That is, the dissimilarity measure $d(x^{(o)},x^{(m)})$ in equation (2.6) is computed from the sample time series to mimic equation (2.5), as if the time series were Gaussian distributed.

Comments: (1) The I-divergence or Kullback Leibler information number (also the information for discrimination, information gain or entropy of f_o relative to f_m) has a basic role in the information theoretic approach to statistics, and in statistical physics as maximization of entropy Kullback, 1968, Good, 1963, Jaynes, 1957. The I-divergence does not satisfy the triangle inequality and is not a metric. Certain analogies do exist between the properties of probability density functions and Euclidean geometry, wherein I-divergence plays the role of squared Euclidean distance, Csiszár, 1975.

(2) In Appendix 1 it is shown that the dissimilarity measure in equation (2.6) has sufficient metric properties for the formal nearest neighbor statistical classification properties to hold. Those properties are that the asymptotic probability of misclassification is bounded between the Bayes risk and twice the Bayes risk, Cover and Hart 1967, and the O(1/N) finite labeled sample cross-validation-leave out one-at-a-time classification of the labeled sample data set to estimate the probability of misclassification, Cover 1969 and Rogers 1977.

(3) The cross validation estimate of the probablity of misclassification permits the implicit conjecture in the exploratory population screening problem investigation, that there is sufficient evidence in the measurement data to achieve statistically satisfactory discrimination, to be tested with only a moderate number of labeled samples.

(4) Another classification problem of interest is the "normalized baseline" time series classification problem. That problem situation is dominated by intrasubject categorical time series variability. An application of nearest neighbor Kullback Leibler type dissimilarity measures to the classification of faults in relating machinery in a normalized baseline classifiation problem context is in Gersch et al 1980b.

3. IMPLEMENTATION AND INTERPRETATION OF THE NEAREST NEIGHBOR TIME SERIES CLASSIFICATION RULE.

The formula for the dissimilarity between the T-duration d-variable sample time series $x^{(o)}$ and $x^{(m)}$ in equation (2.6) indicates operations on matrices of size $Td \times Td$. Almost invariably direct computations on such sized matrices is forbidding. Alternatively, explicit time and frequency domain formula for the specific situation of the Kullback Leibler number between multivariate stationary ergodic Gaussian distributed time series are of interest. Such formulas are developed in Appendix 2. Mimicing those formulas yields practical implementable dissimilarity measure computations that only involve operations on $d \times d$ matrices. A tried and recommended procedure for computing those dissimilarity measures involves the parametric autoregressive (AR) modelling of the $x^{(o)}$ and $x^{(m)}$ time series.

For example, consider the d-variate time series $x^{(j)}$, for $j = 0$ or m and let

$$\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x^{(j)}(x)$$

$$\hat{\Gamma}^{(j)}(k) = \frac{1}{T} \sum_{t=1}^{T-k} (x^{(j)}(t+k) - \bar{x})(x^{(j)}(t) - \bar{x})', \quad k = 0,1,... \qquad (3.1)$$

denote the sample mean and sample covariance of the j-th time series. Then, the autoregressive model of order p_j fitted to $x^{(j)}$ satisfies,

$$\sum_{i=0}^{p_m} \hat{A}^{(j)}(i) x^{(j)}(t-i) = e^{(j)}(t) \ , \quad \hat{A}^{(j)}_{(o)} = I_d$$

$$E[e^{(j)}(t)] = 0, \ E[e^{(j)}(t+k)e^{(j)}(t)'] = \hat{V}_j \delta_{k,0} \ . \qquad (3.2)$$

In equation (3.2) $x^{(j)}(t)$ and $e^{(j)}(t)$ are d-vectors and $\hat{A}^{(j)}(i)$ are $d \times d$ matrices. The AR model in equation (3.1) may be fitted to the labeled sample time series $x^{(m)}$, $m = 1,\ldots,N$ and the new time series $x^{(o)}$ by employing the Whittle-Robinson recursive model computation — Akaike AIC criterion model order selection procedure, Whittle 1963, Akaike 1974. The fitting of multivariate AR models to data and illustrative examples are shown in Akaike 1976 and Gersch and Yonemoto 1977.

Then, a computationally convenient dissimilarity measure between the time series $x^{(o)}$ and $x^{(m)}$ is

$$2d(x^{(o)},x^{(m)}) = \ell n \frac{|\hat{V}_m|}{|\hat{V}_m|} + \mathrm{tr}\Big(\sum_{i=0}^{p_m} \sum_{j=0}^{p_m} \hat{A}^{(m)}(i)\hat{\Gamma}^{(o)}(j-i)\hat{A}^{(m)\prime}(j)\hat{V}_m^{-1}\Big) - d \ . \tag{3.3}$$

Equation (3.3) only involves operations on $d \times d$ matrices. It mimics the second time domain formula in Appendix 2 for the computation of Kullback Leibler numbers between the probability density functions of Gaussian distributed zero-mean stationary time series. The finite duration multivariate time series $x^{(o)}$ and $x^{(m)}$ are modeled by finite order autoregressive models. In equation (3.3), the hatted quantities are estimates of the corresponding theoretical quantities, p_m is the order of the AR modeled time series $x^{(m)}$ and $\hat{\Gamma}^{(o)}(\cdot)$ is the sample covariance matrix function of the new time series $x^{(o)}$.

Figure 1 shows a schematic implementation of the computation of the dissimilarity measure between the new time series $x^{(o)}$ and the labeled sample time series for $m = 1 \ldots, M$. AR models of the $x^{(o)}$ and the labeled sample $x^{(m)}$'s are assumed. Application of the new time series $x^{(o)}(t)$ to the m-th AR model yields the residual time series $e^{(o,m)}(t)$, $t = 1,\ldots,T$. The dissimilarity measure, $d(x^{(o)},x^{(m)})$, can also be expressed in terms of a formula involving the residual time series $e^{(o,o)}(t)$ and $e^{(o,m)}(t)$, Gersch 1977. The term residual is the quantity remaining or not explained after a particular model is fitted to the data. If one of the labeled sample AR time series models is precisely the AR model that corresponds to the generation of the $x^{(o)}(t)$ data, the corresponding residual sequence will be a white noise sequence. In that sense, the nearest neighbor rule selects the "closest to whiteness" residual sequence.

More concisely, the AR models of the labeled time series sample can be interpreted as templates of those time series. In effect, in the nearest neighbor classification procedure, the new time series is compared against the templates of the labeled sample time series. The most similar template is the one for which the dissimilarity measure is smallest.

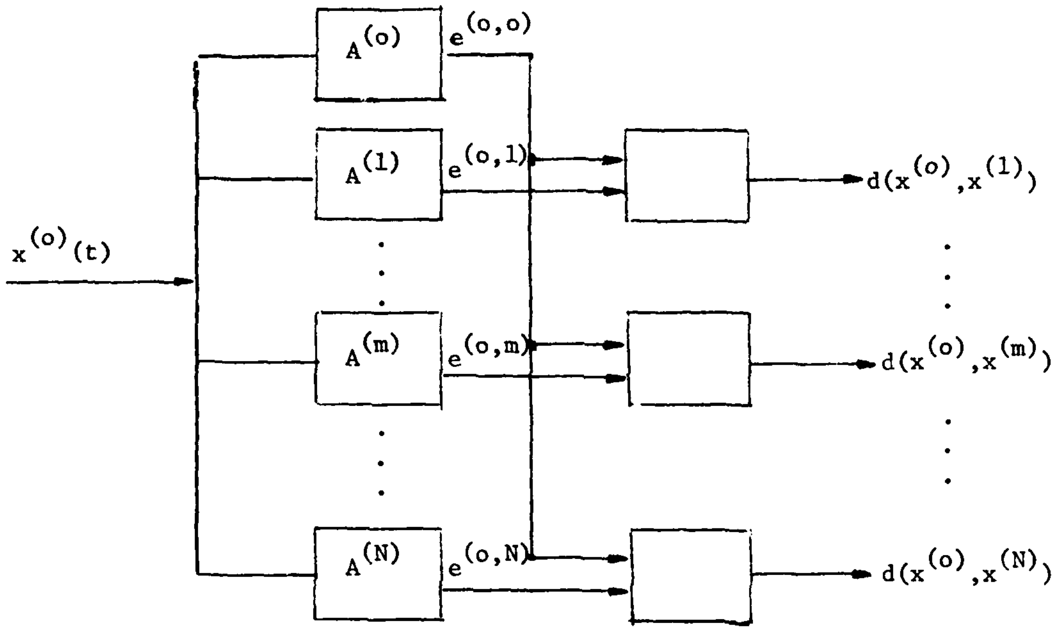

Figure 1. A schematic implementation of a time series data nearest neighbor rule classification procedure.

4. AN ANESTHESIA LEVEL CLASSIFICATION BY EEG ANALYSIS POPULATION SCREENING PROBLEM.

An exploratory EEG time series data-population screening classification problem is treated by the nearest neighbor rule approach. The category or state of an individual is classified by comparison of his or her EEG with EEGs taken from other individuals. The automatic classification of anesthesia levels L1 and L3, respectively the anesthesia levels insufficient for and sufficient for deep surgery by machine computations on the EEG alone is considered. Extension of the nearest neighbor rule approach to distinguish between more than two categories or anesthesia levels does not involve any new concepts.

The anesthesia level EEG data originated in an experiment at Vancouver General Hospital. 280 epochs of visually screened, relatively artifact free, stationary halothane-nitrous oxide anesthesia level labeled EEGs were collected from twenty different individuals in surgery. The non-EEG criteria determined anesthesia levels were classified by a single anesthesiologist to eliminate the problem of inter-EEG-rater variability. Details of the experimental surgical anesthesia situation and a review of the status of automatic classification of anesthesia levels using EEG data appear elsewhere, McEwen, 1975a,b and Gersch et al 1980a. The data consisted of 64 second recordings of four channel EEG epoch data, (F4-C4, F3-C3, C4-02, and C3-01 in the 10-20 EEG system), analogue-FM recorded through a 0.54 to 30 Hz. bandpass filter and subsequently digitially transcribed at the rate of 128 samples/second. An examination of the available data suggested that we confine our attention to a two category classification problem, to classify the anesthesia levels L1 and L3 respectively, the anesthesia levels that are insufficient and just sufficient for deep surgery. The data selected for analysis was the 73 EEG epochs comprised of all the 35-L1 EEG epochs available and 38-L3 EEG data epochs (in sets of 2-3 epochs per individual) from a total of 18 different individuals. The analysis was performed on the first twenty second intervals of each EEG data epoch at a reduced data rate of 128/3 samples per second on $d = 4$ EEG data channel and $d = 2$ EEG data channel (C4-02 and C3-01) data. This constitutes the labeled sample data base.

The implicit conjecture in the EEG population screening problem is that there is sufficient information in the EEG alone to achieve clinically acceptable levels of discrimination between categorical EEG states. The credibility of this conjecture is strained by evidence of the broad intersubject categorical EEG variability. Figure 2, 2-channel twenty second anesthesia level L1 and L3 EEG epochs from five different subjects suggests that the EEG of an individual does differ in the L1 and L3 anesthesia level states and also illustrates broad intersubject EEG variability. The L1 EEGs appear to be relatively homogeneous "fast" EEGs whereas the L3 EEGs include fast, slow, regular and irregular EEGs. No obvious visual properties of the EEGs distinguish the L1 and L3 EEGs from each other.

A useful statement of the conjecture in the EEG population screening problem is: Given labeled EEG samples from two categorical populations, estimate the theoretically best achievable statistical classification performance. The use of the KL number type metric in NN rule classification, in a delete-one subject's EEG-at-a-time KL-NN and KL-kNN classification of the labeled EEG sample base, yields that estimate. (See Duda and Hart 1973 for kNN rules).

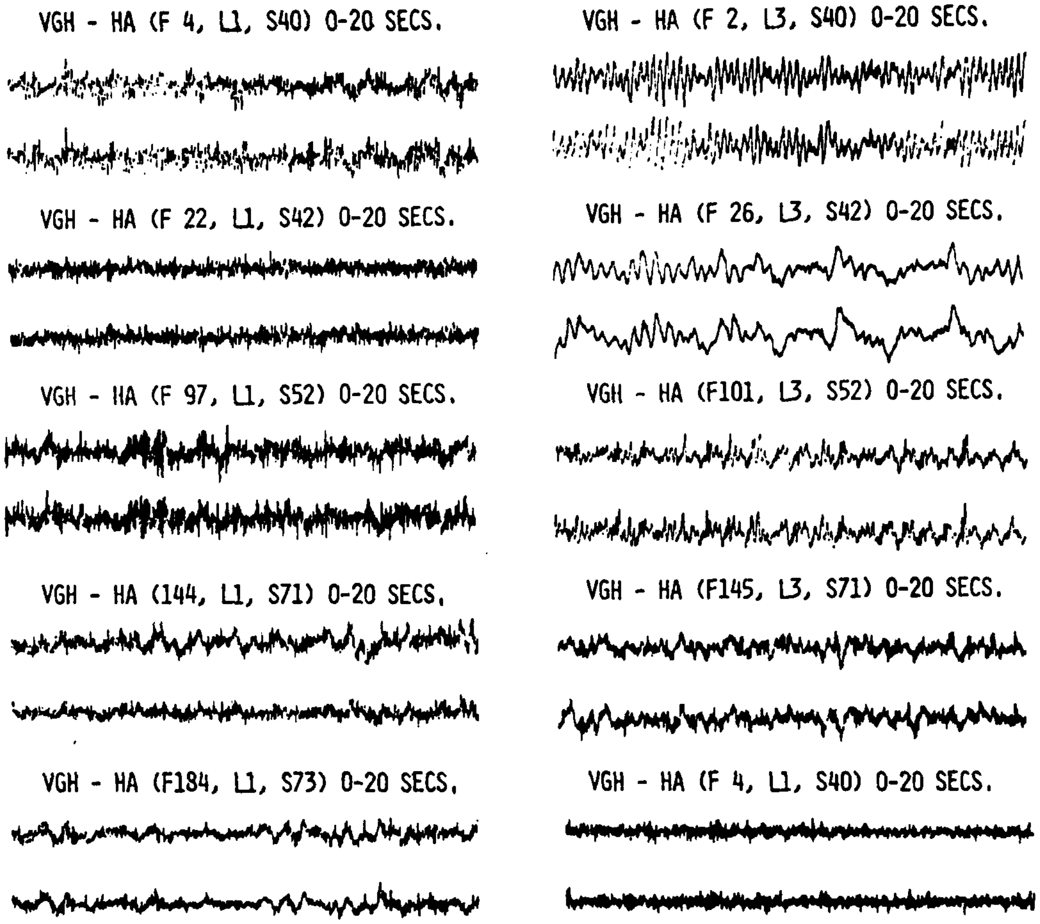

Figure 2. Two channel EEG time series from five different individuals in each of the anesthesia level states L1 and L3.

To achieve a baseline appraisal of the achievable discriminability between the L1 and L3 anesthesia level EEG sample populations, the EEG epochs of a single individual at a time were deleted from the 18 individual - 73 epoch labeled sample EEG data. Each of the deleted-individual's EEG epochs was classified against the remaining 17 individual labeled EEG sample population using KL-NN and KL-kNN rules. The results obtained are shown in Table 1. The entries in the table indicate the number of classification errors and the percentage of correct classification for the best d = 2 EEG channel and d = 4 EEG channel KL-NN classification performance. The best classification results for the d = 2 and d = 4 EEG data channels was 85% and 89% overall correct classification respectively.

TABLE 1: DELETE ONE-SUBJECT-AT-A-TIME, KL-NN RULES RESULTS

KL-3NNN; d = 2

Errors, % Correct Labeled EEGs	L1	L3
L1 35 epochs	1 97%	--
L3 38 epochs	--	10 74%

KL-NN; d = 4

Errors, % Correct Labeled EEGs	L1	L3
L1 35 epochs	1 97%	--
L3 38 epochs	--	7 82%

The objectives of this exploratory population screening anesthesia level classificacion by EEG analysis study have been very clearly met. with only a moderate sized label sample data base, the results obtained quite reliably suggest that the population screening anesthesia level classification by EEG analysis scenario has substantial possibilities for clinical applications.

Comments: 1) Additional considerations for the implementation of nearest neighbor rules in automatic EEG classification such as the consequences of alternative EEG normalizations on classification performance and nearest neighbor data thinnings analysis considerations to economize on computational and storage burdens are examined in Gersch et al 1980a. Briefly, any comparison of EEG time series is subject to arbitrary conventions, criteria and normalizations. The alternative normalizations of the EEG that are possible in nearest neighbor rule classification are explicit in equation (3.3), the dissimilarity measure formula between stationary time series EEGs. The alternative normali..ations influence the relative dominance of the first and second terms in that equation. The related subject of distortion measures for speech processing is treated by Gray et al 1980.

2) Time series classification by nearest neighbor rules with Kullback Leibler type dissimilarity measures for classification are very well known in speecsing, Itakura and Saito, 1970, and Gray et al 1980. In those applications KL number dissimilarity measures most commonly involve the modelling of each of the labeled sample and new (scalar) time series by fixed order autoregressive models. Those AR model parameters or features are transformed into the Kullback Leibler number measures. Because the order of the AR models fitted to each time series is fixed, that classification procedure is potentially of the feature analysis-discriminant analysis variety. The poignant remark by Cover 1973, that the problem for which that solution is optimum is not known is applicable here. Thus the usual speech processing adaption of NN-KL type metric classification has no necessary statistical

nor nearly optimal statistical classifiation properties. An example of the misclassification of time series that results from arbitrarily fixing the order of AR model fitted to the time series is in Brotherton and Gersch, 1980.

APPENDICES

1. THE METRIC PROPERTIES OF $d(x^{(o)},x^{(m)})$.

Here we show that the Kullback Leibler type dissimilarity measure between time series has sufficient metric properties that the formal nearest neighbor rule statistical classification properties apply to nearest neighbor classification rules that employ that measure. Following the development of Cover and Hart 1973, it is only necessary to show that i) $d(x^{(o)},x^{(o)}) = 0$, ii) $d(x^{(o)},x^{(m)}) > 0$ for any $x^{(m)} \neq x^{(o)}$, and iii) the minimum value of the dissimilarity measure $d(x^{(o)},x^{(m)}) \to 0$, as N the number of labeled samples increases indefinitely. Property i) is immediate from equation (2.6). Property ii) is proved below separately. To prove property iii); Let the sample $Td \times Td$ covariance matrix $\hat{\Sigma}_o$ be distributed in accordance with distribution F. Let $\hat{\Sigma}_{o,T}, \hat{\Sigma}_{1,T}, \hat{\Sigma}_{2,T},\ldots$ be IID random variables from that distribution. Then the space $R^{Td \times Td}$ on which the sample covariances are defined is a separable metric space and the minimum Euclidean distance between the sample covariances goes to zero. That is, $||\hat{\Sigma}_o - \hat{\Sigma}_{m'}|| \to 0$. Then, since $d(x^{(o)},x^{(m)}) = d(\hat{\Sigma}_o,\hat{\Sigma}_{m'})$ is a continuous function of $\hat{\Sigma}_{m'}$ and $\hat{\Sigma}_{m'} \to \hat{\Sigma}_{o,T}$ in $R^{Td \times Td}$, $d(\hat{\Sigma}_{o,T},\hat{\Sigma}_{m',T}) \to 0$, Royden, 1968. Property ii); Consider the situation with $\hat{\Sigma}_o = \hat{\Sigma}_o$ and $\hat{\Sigma}_m = \hat{\Sigma}_o + \Delta$. The matrix Δ denotes a small perturbation matrix. For convenience subscript T and hat notation will be dropped in what follows. Then

$$2Td(x^{(o)},x^{(m)}) = 2Td(\Sigma_o,\Sigma_o+\Delta) = [\ell n \frac{|\Sigma_o+\Delta|}{|\Sigma_o|} + tr(\Sigma_o(\Sigma_o+\Delta)^{-1})-Td] .$$

We would like to prove that $d(\Sigma_o, \Sigma_o + \Delta) \geq 0$. Σ_o is symmetric, positive definite and fixed, $\Sigma_o + \Delta$ symmetric and positive definite and Δ is "small". Let $A = \Sigma_o^{-1/2} \Delta \Sigma_o^{-1/2}$. Then A is symmetric, and $I + A$ is positive definite. Then, the last equation can be written

$$f(A) = \ell n|I+A| + tr(I+A)^{-1} - Td \ .$$

The problem is now reduced to demonstrating that $f(A)$ is convex in the neighborhood of $A + 0$. Let $A = A(s)$ be linear in s. Then by the rules, $d\ell n|X| = tr(X^{-1})(dX)$ and $dX^{-1} = -(X^{-1})(dX)(X^{-1})$, Anderson, 1958,

$$\frac{d\ f(A)}{ds} = tr(I+A)^{-1} (\frac{dA}{ds}) - tr(I+A)^{-1} (\frac{dA}{ds})(I+A)^{-1}$$

where $(\frac{dA}{ds})$ is a symmetric matrix. Since $A(s)$ is linear, $\frac{d^2A(s)}{ds^2} = 0$, so

$$\frac{d^2f(A)}{ds^2} = - tr(I+A)^{-1} (\frac{dA}{ds})(I+A)^{-1}(\frac{dA}{ds})$$

$$+ 2\ tr(I+A)^{-1} (\frac{dA}{ds})(I+A)^{-1} (\frac{dA}{ds})(I+A)^{-1}$$

$$= tr(I+A)^{-1} (\frac{dA}{ds}) [2(I+A)^{-1} - I] (\frac{dA}{ds})(I+A)^{-1}$$

$$= tr\ B[2(I+A)^{-1} - I]B' \ .$$

In the equation above we have let $B = (I+A)^{-1}(\frac{dA}{ds})$. The right-hand side of that equation will be non-negative provided the term in brackets in the last row is positive semidefinite. But that is equivalent to $2I - (I+A)$, and the $(I-A)$ being positive semidefinite. That implies $A \leq 1$ in the sense of positive definiteness and also $-I \leq A$ since $I+A \geq 0$. Then, clearly when $A = 0$, $\frac{d^2f(A(s))}{ds^2} > 0$ and in general $f(A)$ is convex provided $-I < A < I$.

A2. TIME AND FREQUENCY DOMAIN KULLBACK LEIBLER FORMULAS BETWEEN STATIONARY GAUSSIAN TIME SERIES.

TIME SERIES REPRESENTATIONS. Let $\{x^{(o)}(t)\}$ and $\{x^{(m)}(t)\}$ denote d-variabe zero-mean stationary ergodic Gaussian time series with corresponding probability density function $f^{(o)}$, $f^{(m)}$ and $d \times d$ matrix covariance functions $\Gamma^{(o)}(k)$, $\Gamma^{(m)}(k)$ and power spectral density matrices $S_o(f)$ and $S_m(f)$ respectively. Identify the time series $x^{(i)}(t)$ parametrically in terms of the Wold (moving average) and autoregressive representations, Whittle, 1963.

$$x^{(i)}(t) = h^{(i)}(t) * \varepsilon^{(i)}(t) \qquad i = 0,1,2,\dots,M$$

$$A^{(i)}(t) * x^{(i)}(t) = \varepsilon^{(i)}(t);$$

$$E(\varepsilon^{(i)}(t)) = 0; \quad E(\varepsilon^{(i)}(t+k)\varepsilon^{(i)}(t)') = V_i \, \delta_{k,0} \qquad \text{(A1)}$$

In equation (A1), the symbol $*$ denotes the convolution operation, E is the expectation operator and $\{h^{(i)}(t)\}$ and $\{A^{(i)}(t)\}$ are respectively the $d \times d$ impulse matrix response and AR matrix coefficients. Denote the action of the AR operator defined by $x^{(m)}(t)$ on $x^{(o)}(t)$ by

$$A^{(m)}(t) * x^{(o)}(t) = e^{(o,m)}(t) . \qquad (A2)$$

In equation (A2), $e^{(o,m)}(t)$ has an interpretation as a zero-mean "residual" time series in the conventional sense of a regression analysis. Its zero-log covariance matrix is

$$E(e^{(o,m)}(t)e^{(o,m)}(t)') \equiv V_m^o . \qquad (A3)$$

Employing the notation of equation (A2) in equation (A1)

$$\begin{aligned} A^{(m)}(t) * h^{(o)}(t) * \varepsilon^{(o)}(t) &= e^{(o,m)}(t) \\ h^{(o,m)}(t) * \varepsilon^{(o)}(t) &= e^{(o,m)}(t) . \end{aligned} \qquad (A4)$$

In equation (A4), $h^{(o,m)}(t)$ designates the impulse response of the cascade of filters $A^{(m)}(t)$ and $h^{(o)}(t)$. By elementary linear operations,

$$h^{(o,m)}(t) = h^{(o)}(t) + \sum_{i=1}^{\infty} A^{(m)}(i)h^{(o)}(t-i), \ t = 0,1,... \qquad (A5)$$

KULLBACK LEIBLER NUMBER FORMULAS. Then, time and frequency domain formulas for the Kullback Leibler numbers between those Gaussian time series are:

$$\begin{aligned} 2I(f^{(o)},f^{(m)}) &= \ell n \frac{|V_m|}{|V_o|} + tr(\sum_{t=0}^{\infty} h^{(o,m)}(t) V h^{(o,m)'}(t) V_m^{-1}) - d \\ &= \ell n \frac{|V_m|}{|V_o|} + tr(\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A^{(m)}(i) \Gamma^{(o)}(j-i) A^{(m)'}(j) V_j^{-1}) - d \\ &= \ell n \frac{|V_m|}{|V_o|} + tr(\int_{-\frac{1}{2}}^{\frac{1}{2}} S_o(f) S_m(f)^{-1} \, df) - d \ . \qquad (A6) \end{aligned}$$

An intermediate result, derived from equation (2.6) and the time series notation above, from which the results in equation (A6) follow is that

$$2I(f^{(o)},f^{(m)}) = \ell n \frac{|V_m|}{|V_o|} + tr[V_m^o \, V_m^{-1}] - d \ . \qquad (A7)$$

Then, the first two parametric time domain formulas for the Kullback Leibler number between stationary Gaussian time series in equation (A6) may be derived from equation (A7) by replacing V_m by its definition, equation (A3) and then substituting for $e^{(o,m)}$ by its representations in equations (A4) and (A2) and taking the indicated expectations. The frequency domain formula, the third line in equation (A6) is obtained from equation (A7) by the use of Parseual's theorem and the assumption of ergodicity.

Comments: The first development of a frequency domain formula for the Kullback Leibler number between Gaussian distributed time series was probably due to Pinsker, 1964, that work appears to have remained almost unknown to Western researchers. Subsequently frequency domain formulas were developed by Shumway and Unger, 1974, Hawkes and Moore, 1976 and

B.D.O. Anderson et al 1978. The first time domain formula for the Kullback Leibler number between scalar time series is probably due to Itakura and Saito 1968, in their search for distance measures in speech classification. A complete and up-to-date treatment of that approach is in Gray et al 1980. Akaike, 1976 shows different development of the second time domain formula in equation (A6). That development has attracted little attention.

Will Gersch
Dept. of Information &
Computer Science
University of Hawaii
Honolulu, Hawaii 96822

BIBLIOGRAPHY

Akaike, H., "A new look at the statistical model identification," IEEE Transactions on Automatic Control, AC-19, 716-722, 1974.

Akaike, H., "Canonical correlation analysis of time series and the use of an information criterion," in Mehra, R.K. and Lainotis, D.G. (eds.), System Identification: Advances and Case Studies, Academic Press, New York, 27-97, 1976.

Anderson, T.W., Introduction to Multivariate Analysis, John Wiley, New York, 1958.

Brotherton, T. and Gersch, W. Nearest neighbor and parametric AR model classification of stationary time series, in preparation.

Cover, T.M., Learning in pattern recognition. In Methodologies of Pattern Recognition, Ed. S. Watanabe, Academic Press, 1969.

Cover, T.M., Recent books on pattern recognition; A review IEEE Trans. Information Theory, IT-19, 827-833, 1973.

Cover, T.M. and Hart, P.E., "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, IT-13, 21-27, 1967.

Csiszar, I., "I-divergence geometry of probability distributions and minimization problems," The Annals of Probability, Vol. 3, No. 1, 146-158.

Duda, R.O. and Hart, D.E., Pattern Classification and Scene Analysis, Wiley, New York, 1973.

Gersch, W., "Discrimination Between Stationary Gaussian Processes, Large Sample Results," TR No. 30, Dept. of Statistics, Stanford University, 1977 T.W. Anderson, Project Director.

Gersch, W., Brotherton, T. and Braun, S., Nearest neighbor-time series analysis classification of faults in rotating machinery, Proc. Int'l Conf. on Reliability, Stress Analysis & Failure Prevention (ASME), August 1980b, San Francisco.

Gersch, W., Martinelli, F., Yonemoto, J., Low, M.D. and EcEwen, J.A., "Automatic Classification of Electroencephalograms: Kullback Leibler-Nearest Neighbor Rules," Science, 205, 193-195, 1979.

Gersch, W., Martinelli, F., Yonemoto, J., Low, M.D. and McEwen, J.A., "Kullback Leibler-Nearest Neighbor Rule Classification of EEGs: The EEG Population Screening Problem, and Anesthesia Level EEG Classification Application," Computers and Biomedical Research, 16, 1980a, in press.

Gersch, W., and Yonemoto, J.,"Parametric time series models for multivariate EEG analysis," Computers and Biomedical Research, 10, 113-125, 1977.

Gevins, W., "Application of pattern recognition to brain potentials and electrical potentials," IEEE Trans.on Pattern Recognition and Machine Intelligence PAMI 2, xxx-xxx, 1980, in press.

Good, I.J., "Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables," Ann. Math. Statist., 34, 911-934, 1963.

Gray, R.M., Buzo, A., Gray Jr., A.H. and Matsuyama, Y., "Distortion measures for speech processing," IEEE Trans. ASSP, in press.

Itakura, F., and Saito, S., "Analysis synthesis telephony based upon the maximum likelihood method," Report of the 6th International Congress of Acoustics, Y. Kohas, ed., Tokyo C-5-5, C-17-20, 1968.

Jaynes, E.G., "Information theory and statistical mechanics," Phys. Rev., 106, 620-630, 1957.

Kullback, S., Information Theory and Statistics, Wiley, New York, 1958, Dover, New York, 2nd edition, 1968.

McEwen, J.A., "Estimation of the level of anesthesia during surgery by automatic EEG pattern recognition, Ph.D. Thesis, British Columbia, Vancouver, 1975.

McEwen, J.A., Anderson, G.B., Low, M.D., and Jenkins, L.C., "Monitoring the level of anesthessia by automatic analysis of spontaneous EEG activity," IEEE Trans. Bio-Med. Eng., 22, 299-305, 1975.

Rogers, W.H., "Some convergence Properties of k-nearest Neighbor Estimates, " Ph.D. Thesis, Statistics Department, Stanford University, April 1976.

Royden, H.L., Real Analysis, Macmillian, 1968.

Whittle, P., "On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix," Biometrika, 60, 129-134, 1963a.

Whittle, P., Prediction and Regulation, Van Nostrand, 1963b.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
294

2. GOVT ACCESSION NO.
AD A096 952

3. RECIPIENT'S CATALOG NUMBER

4. TITLE *(and Subtitle)*
NEAREST NEIGHBOR CLASSIFICATION OF STATIONARY TIME SERIES: AN APPLICATION TO ANESTHESIA LEVEL CLASSIFICATION BY EEG ANALYSIS

5. TYPE OF REPORT & PERIOD COVERED
TECHNICAL REPORT

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
WILL GERSCH

8. CONTRACT OR GRANT NUMBER(s)
N00014-76-C-0475

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
Stanford University
Stanford, CA 94305

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR-042-267

11. CONTROLLING OFFICE NAME AND ADDRESS
OFFICE Of Naval Research
Statistics & Probability Program Code 436
Arlington, VA 22217

12. REPORT DATE
DECEMBER 5, 1980

13. NUMBER OF PAGES
25

14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*

15. SECURITY CLASS. *(of this report)*
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Time series classification; nearest neighbor rules; Kullback Leibler numbers; EEG classification.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

PLEASE SEE REVERSE SIDE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

F

294

An exploratory time series data-population screening problem is considered. A moderate, but not a large, number of categorical or labeled time series are obtained from different individuals (or objects). There is broad intersubject time series variability in each category. The objectives are to obtain a statistically reliable estimate of the minimum achievable probability of misclassification of new time series and to implement a time series classification rule that can achieve that statistical performance.

A nearest neighbor classification rule achieves those objectives. With that rule a measure of dissimilarity is computed between a new to-be-classified time series and each of a set of categorically labeled time series. The new time series is classified with the label of its least dissimilar neighbor. In our approach the dissimilarity measure between the time series is an estimate of the Kullback Leibler number between the time series computed as if the time series were normally distributed. This dissimilarity measure is shown to have sufficient metric properties for the formal Cover and Hart asymptotic nearest neighbor and Rogers finite sample nearest neighbor classification rule properties to hold.

Additional results include time and spectral domain formulas for the Kullback Leibler numbers between stationary Gaussian distributed time series and a practical computational method for the estimation of Kullback Leibler type dissimilarity numbers between time series. The nearest neighbor Kullback Leibler type dissimilarity measure classification rule method is applied to the classification of the level of anesthesia of humans in surgery by the analysis of electroencephelograms.

B